DOCUMENT RESUME

ED 296 577 FL 017 457

AUTHOR Perkins, Kyle; Duncan, Ann
TITLE An Item Discriminability Study of the Iowa Tests of Basic Skills (Form G, Level 12) Language Skills Test.
PUB DATE [88]
NOTE 26p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Correlation; Item Analysis; *Language Skills; *Language Tests; Language Usage; Punctuation; Spelling; Standardized Tests; Test Construction; *Test Items
IDENTIFIERS *Iowa Tests of Basic Skills; *Item Discrimination (Tests)

ABSTRACT

An item discriminability study of the Iowa Tests of Basic Skills Language Skills tests identified test items that are robust discriminators, psychometrically capable of separating low scorers from higher scorers in the language tests battery. The analysis was conducted by calculating a point-biserial correlation for each item on the four language skills subtests in spelling, capitalization, punctuation, and usage and expression. Results show the spelling test to be the most robust in terms of having the highest average point-biserial correlation and the highest total percentage of items meeting the discriminability criterion. The other three subtests clustered together far below the spelling test. The findings are intended for the use of teachers and administrators using this and other standardized tests for the same skills objectives, and teachers and test developers interested in developing their own assessment instruments. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED296577

An Item Discriminability Study of the Iowa Tests of Basic Skills (Form G, Level 12) Language Skills Tests

Kyle Perkins
Professor
Department of Linguistics
Southern Illinois University
Carbondale, Illinois 62901-4517

and

Ann Duncan
Principal
Carterville Grade School
Carterville, Illinois 62918

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

L. Perkins

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

✓ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

FL017457

Abstract

For each item from the Iowa Tests of Basic Skills Language Skills test a point-biserial correlation was computed and corrected for part-whole overlap. The criterion for acceptable item discriminability was set at .40.

Nineteen of the forty-one Spelling items exhibited acceptable item discriminability indices. Eight of the thirty Capitalization items yielded point-biserial correlations of .40 or higher. Eleven of the forty Usage and Expression items were acceptable discriminators. And seven of the thirty Punctuation items produced acceptable item discriminability indices.

The purpose of the research was to determine which skills objectives from the Iowa Tests of Basic Skills Language Skills test were good discriminators. The intended audience includes teachers and administrators who use the Iowa battery, those who use other standardized tests which assess the same skills objectives as the Iowa battery measures, and teachers and test developers who have interest in developing their own assessment instruments.

An Item Discriminability Study of the Iowa Tests of Basic Skills (Form G, Level 12) Language Skills Tests

INTRODUCTION

This paper reports the results of an item discriminability study of the Iowa Tests of Basic Skills (Form G, Level 12) Language Skills tests (Hieronymus, Hoover, and Lindquist, 1986). Our purpose for conducting this research was to identify the items in the Iowa battery which were robust discriminators, items which were psychometrically robust to the extent that they separated the lower scorers from the higher scorers in the language tests battery. This body of information should prove useful to those teachers and administrators who use the Iowa battery, to those who use standardized tests other than the Iowa tests but assess the same skills objectives as the Iowa tests, and to those teachers and test developers who are considering the development of their own classroom and achievement tests. The results reported in this paper should provide useful guidelines to test developers.

Item discriminability

Before we describe the study and its results, we want to establish the importance of item discriminability. Item difficulty and item discriminability are two important concepts in item analysis and test evaluation. Item difficulty is determined as the proportion of correct responses. Tuckman (1978) has argued that items with a proportion of correct responses less than .33 should be rejected as well as items with a proportion of correct responses greater than .67. Items with an item difficulty greater

than .67 are considered to be too easy, while items with an item difficulty less than .33 are considered to be too difficult. However, many test developers will include items at the extreme poles of the difficulty continuum in order to insure content validity, to shape the information curve, and to help the examinee to overcome psychological inertia.

Item discriminability has to do with how well an item discriminates between low scorers (weak examinees) and high scorers (strong examinees) on the ability or trait being tested.

Henning (1987) has motivated the importance of item discriminability in test evaluation:

> Difficulty alone is not sufficient information upon which to base the decision ultimately to accept or reject a given item. Consider, for example, a situation in which half of the examinees pass a given item and half fail it. Using difficulty as the sole criterion, we would adopt this item as an ideal item. But what if we discovered that the persons who passed the item were the weaker half of the examinees, and the persons who failed the item were the stronger examinees in the ability being measured. This would certainly cause us to have different thoughts about the suitability of such an item. If our test were comprised entirely of such items, a high score would be an indication of inability and a low score would be an indication of comparative ability (p. 51).

METHODOLOGY

Subjects

Eighty-two sixth-grade students enrolled at Carterville (Illinois) Grade School generated the data for this research. The data were collected from the Iowa Tests of Basic Skills Form G, Level 12 battery (Hieronymus, Hoover, and Lindquist, 1986) which is administered during the spring of each school year as a part of the school's assessment program. We do not have demographic data on these students because the item responses were coded anonymously.

Instrumentation

The Iowa Language Skills battery consists of four tests: (1) Spelling; (2) Capitalization; (3) Punctuation; and (4) Usage and Expression. The following descriptions of the tests are found in the Preliminary Teacher's Guide (Hieronymus, Hoover, and Lindquist, 1986).

Spelling

The Form G, Level 12 Spelling test contains forty-one items; working time was twelve minutes.

> The item classification system for Test L-1: Spelling is based upon error types. For each item, the type of error is indicated for the one word in four (if any) that is misspelled. Of course, pupils may in fact commit a double error in overlooking a misspelled word and in marking as wrong a word that is correctly spelled.
>
> Subskills are coded as 25 consonant phonemes, 7 simple vowels,

and 13 compound vowel nuclei. The coding used in each skills category is as follows:

CS Consonant Substitutions: the consonant sound in which the error occurs is listed (four items).

CR Consonant Reversals: reversals of consonants and vowels were listed as consonant reversals; the correct order is listed (four items).

CO Omission of Consonants: the letter or letters omitted are listed (four items).

CA Addition of Unnecessary Consonants: the letter or letters added are listed (five items).

CD Double Consonants-a special case of omission or addition: a is used to indicate that the second consonant has been added where it does not belong; o is used to indicate omission of one of the consonants (four items).

VS Vowel Substitution: the vowel sound in which the error occurs is listed (four items).

VR Vowel Reversals: the correct order of vowels is listed (four items).

VO Omission of Vowels: the letter omitted is listed (four items).

VA Addition of Unnecessary Vowels: the letter added is listed (three items) (pp. 48-49).

Capitalization

The Form G, Level 12 Capitalization test contains thirty items; working

time was twelve minutes. The skills objectives were grouped into the following major categories: (1) names and titles (five items); (2) dates and holidays (one item); (3) place names (seven items); (4) organizations and groups (seven items); (5) linguistic conventions (four items): (6) overcapitalization (three items); and (7) no mistakes (three items).

Punctuation

The Form G, Level 12 Punctuation test contains thirty items; working time was fourteen minutes. The skills objectives which were assessed include the following: (1) use of terminal punctuation (seven items); (2) use of comma (six items); (3) other punctuation: use of apostrophe, quotation marks, colon, and semicolon (eight items); (4) overpunctuation (six items); and (5) no mistakes (three items).

Usage and Expression

The Form G, Level 12 Usage and Expression test contains forty items and assesses twenty-eight skills objectives; working time was sixty-five minutes. The twenty-eight skills objectives are classified into seven major categories: (1) use of verbs (six items); (2) use of personal pronouns (two items); (3) use of modifiers (four items); (4) use of context (seven items); (5) no mistakes (three items); (6) conciseness and clarity (three items); (7) appropriateness (eight items); and (8) organization (seven items).

Analysis

Point-biserial correlations

For each test item we calculated a point-biserial correlation, an item-test total correlation coefficient. The point-biserial correlation is a commonly utilized technique for the computation of item discriminability. Because such correlations are artificially inflated due to the fact that an item is a component of the test score total, we corrected each point-biserial correlation for part-whole overlap thereby removing the contribution of the item score to the total score in a post-hoc fashion.

Following Mehrens and Lehmann (1984), we set the criterion for acceptable discriminability at .40. In other words, if an item had a point-biserial correlation corrected for part-whole overlap of .40 or higher, we considered it to be a robust discriminator. If an item yielded a point-biserial correlation corrected for part-whole overlap of .39 or lower, we considered that item to lack sufficient discriminability robustness.

RESULTS

(Put Table 1 about here)

Table 1 presents the means, standard deviations, grade equivalencies, ranges, and estimates of reliability for the four tests. The Kuder-Richardson Formula 20 was used to estimate the internal consistency reliability of the measures.

(Put Figures 1, 2, 3, and 4 about here)

Figures 1-4 present the frequency distributions for the measures. As the figures indicate, there is some evidence of negative skew, that is, many of the student scores are clustered toward the high end of the scoring continuum. With regard to the school's ability to instill the skills objectives, as measured by the Iowa battery, this does not constitute a problem, for this type of distribution is indeed desirable as an indication that many, if not a majority, of the students achieved the objectives of the instructional program.

On the other hand, skewed distributions can cause some psychometric problems. A characteristic of skewed distributions is that scores tend to be compacted at one end of the scoring continuum, in this case, toward the high side. This phenomenon is referred to as boundary effect, and the result is that one has difficulty in distinguishing among students in the ability at the compacted end of the scoring continuum, resulting in a loss of person separability and reliability.

Spelling

Nineteen of the forty-one Spelling items exhibited point-biserial correlations corrected for part-whole overlap of .40 or higher. The types of errors for the one word that was misspelled in each item can be attributed to the lack of a one-to-one correspondance between English graphemes and their corresponding phonemes. A grapheme is an allograph which represents a single letter of the alphabet, while a phoneme is a class of similar but articulatorily different sounds which are heard as the same sound by native speakers and are represented by the same symbol in phonemic transcription.

One of the four Consonant Substitution items met the discriminability

criterion. The item involved the substitution of the grapheme k for t in a word-final consonant cluster, /kt/.

One of the four Consonant Reversal items met the criterion; the item involved the reversal of a consonant and a vowel. The word-final consonant cluster was /bl/. The examinees probably chose the grapheme combination b e l instead of b l e because these graphemes are pronounced as /b/ followed by a syllabic /l̩/.

Three of the four Omission of Consonants items met the criterion. One item involved a word-final mn grapheme cluster (the n was deleted) which is pronounced /m/; a second, a word-medial cluster ck which is pronounced /k/ (the c was deleted); the third item involved a word-medial br cluster which is pronounced /b/ (the r was deleted).

All five of the Addition of Unnecessary Consonants items met the criterion. An s was added to a word-medial xpr grapheme cluster; the item had to do with the grapheme x being pronounced as /ks/, and the /k/ and /s/ were separated by a syllable boundary. An h was added to a word-medial ss grapheme cluster because the grapheme cluster is pronounced as an /s̆/ or /sh/ depending on the transcription system emplyed. An h was added to a single word-medial s grapheme because the intervocalic s is pronounced /z̆/ or /zh/ due to a contiguous voicing environment. The grapheme k was added to the grapheme c in a word-medial position. In the environment cited, the grapheme c is pronounced as a /k/. The grapheme s was added to the grapheme x; in the environment cited, the grapheme x is pronounced as /ks̆/ or /ksh/, and the two phonemes are separated by a syllable boundary.

Only one of the four Double Consonants items met the criterion: an s grapheme was doubled, and it is pronounced as an /s/.

Not one of the four Vowel Substitution items met the criterion.

Two of the four Vowel Reversal items met the criterion: an ea cluster was reversed, and an ou cluster was reversed. The ea cluster is pronounced as /ɛ/; the ou cluster, /ə/.

All four Omission of Vowel items met the criterion. The o was deleted from an ou grapheme cluster (pronounced as /u/). The i from an ia grapheme cluster was deleted (pronounced as /I/). The e was deleted from a d e r collocation because, in the environment cited, these graphemes are pronounced as /d/ followed by a vocalic /ɹ̩/. The grapheme a was deleted from a ua grapheme cluster; in the cited environment, the ua grapheme cluster is pronounced as /u:/.

One of the three Addition of Unnecessary Vowel items met the criterion: the grapheme a was added after the grapheme o; in the context cited, the grapheme o is pronounced /o:/.

One of the five No Mistakes items met the criterion.

As classes of skills objectives assessed by the Spelling test, Omission of Consonants, Addition of Unnecessary Consonants, and Omission of Vowels were the most robust classes of items vis-a-vis acceptable indices of item discriminability.

Capitalization

Eight of the Capitalization items yielded acceptable indices of discriminability.

The most robust class of items with regard to acceptable discriminability was Overcapitalization: all items met the criterion.

The second most productive class of items was Place Names, with three of the seven items meeting the criterion.

One Names and Titles item and one Organizations and Groups item met the criterion; the sole Dates and Holidays item did not meet the criterion.

Punctuation

Seven of the Punctuation items met the criterion for acceptable discriminability.

Two of the seven Terminal Punctuation items met the criterion: the exclamation mark and the question mark.

Two of the eight Other Punctuation items yielded acceptable discriminability indices: the use of quotation marks before and after a direct quotation and with titles.

Two of the six Overpunctuation items met the criterion: the use of quotation marks with indirect quotations and the use of the comma to mark a trivial pause.

One of the six Use of Comma items met the criterion: to separate the names of city and state; not one of the No Mistakes items met the criterion.

Usage and Expression

Eleven of the forty Usage and Expression items produced point-biserial correlations corrected for part-whole overlap of .40 or higher. Two of the Use of Verbs items met the criterion, both assessing the concord of subject and verb.

Not one of the two Use of Personal Pronouns items produced acceptable discriminability indices.

Two of the four Use of Modifiers items met the criterion: one item assessed the correct comparative form of a monosyllabic adjective; the second item assessed the past participle form of a verb which functioned as an attributive adjective.

Three of the seven Use of Context items produced acceptable discriminability indices. Two of these items involved double negatives; the third assessed the redundant, incorrect use of two root modal auxiliary verbs.

The three No Mistakes items did not meet the criterion.

One of the three Conciseness and Clarity items produced an acceptable discriminability index; the item assessed the use of a dangling present participle modifier.

Two of the eight Appropriateness items met the criterion of a .40 or higher point-biserial correlation. One item assessed the use of an embedded infinitive "predicate", and the second item assessed the use of parallel coordinate structures as the objects of a preposition.

Only one of the seven Organization items met the criterion: the item assessed the correct word order of the principal verb, the direct object, a manner adverbial, and a locative adverbial in a sentence.

In terms of raw percentages of items from each class which met the criterion of discriminability, the skills objectives are rank ordered as follows: first, Use of Modifiers, 50 percent; second, Use of Context, 42.85 percent; tied at third (3.5), Use of Verbs and Conciseness and Clarity, 33.33 percent; fifth, Appropriateness, 25 percent; sixth, Organization, 14,28 percent; and tied at seventh (7.5), Use of Personal Pronouns and No Mistakes, 0 percent.

CONCLUSION

We had a very practical goal in mind when we conducted this research: we wanted to discover which skills objectives from the Iowa Tests of Basic Skills (Form G, Level 12) Language Skills tests were good discriminators; that is, we wanted to discover which items from the Language Skills battery did a good job of separating the weak examinees from the strong examinees. To do this, we calculated a point-biserial correlation for each item and then corrected (reduced) the correlation coefficient for part-whole overlap. We then selected the items which had point-biserial correlations of .40 and higher and reported what those items assessed.

Our intended audience for this research includes teachers and administrators who use the Iowa battery, teachers and administrators who use other standardized tests which assess the same skills objectives as the Iowa battery measures, and teachers and test developers who have interest in developing their own assessment instruments.

(Put Table 2 about here)

In terms of a useful summary of this research, Table 2 presents the range of the point-biserial correlations corrected for part-whole overlap for each test, the average point-biserial correlation corrected for part-whole overlap for each test, and the percentage of items from each test which yielded a point-biserial correlation corrected for part-whole overlap of .40 or higher. We used the Fisher Z transformation to transform the point-biserial correlations from a distorted ordinal scale to a normal interval scale; then we averaged the Fisner Z scores, and transformed the average Z score back to a point-biserial correlation coefficient.

The data in Table 2 indicate that as subtests in the Language Skills

battery, the Spelling test was the most robust, both in terms of having the highest average point-biserial correlation and the highest total percentage of items which met the discriminability criterion. For both these parameters, the Capitalization, Punctuation, and Usage and Expression tests tend to cluster together, far lower than the Spelling test. An interesting avenue for future research would be to determine why this was the case.

REFERENCES

Henning, G. H. A Guide to Language Testing. Development, Evaluation, Research. Cambridge, MA: Newbury House Publishers, 1987.

Hieronymus, A. N., Hoover, H. D., and Lindquist, E. F. Iowa Tests of Basic Skills, Form G, Level 12. Chicago: The Riverside Publishing Company, 1986.

Hieronymus, A. N., Hoover, H. D., and Lindquist, E. F. Preliminary Teacher's Guide. Chicago: The Riverside Publishing Company, 1986.

Mehrens, W. A., and Lehmann, I. J. Measurement and Evaluation in Education and Psychology, third edition. New York: Holt, Rinehart and Winston, 1984.

Tuckman, B. W. Conducting Educational Research, second edition. New York: Harcourt Brace Jovanovich, 1978.

Table 1
Means, standard deviations, and estimates of reliability

Test	No. of students	No. of items	Mean	Grade equivalency	S.D.	Range	Estimate of reliability
Spelling	82	41	27.59	76	7.40	8 to 40	.885
Capitalization	82	30	22.22	78	4.64	8 to 29	.793
Punctuation	82	30	21.23	76	4.79	6 to 30	.789
Usage and Expression	82	40	30.01	79	5.76	15 to 39	.816

Table 2
Range of point-biserial correlations and percentage of items which were robust discriminators

Test	Range	Average point-biserial correlation	Robust discriminators: percentage of items
Spelling	-.011 to .596	.364	46.34
Capitalization	-.009 to .708	.320	26.67
Punctuation	.011 to .565	.306	23.33
Usage and Expression	-.159 to .667	.304	27.50

Figure 1

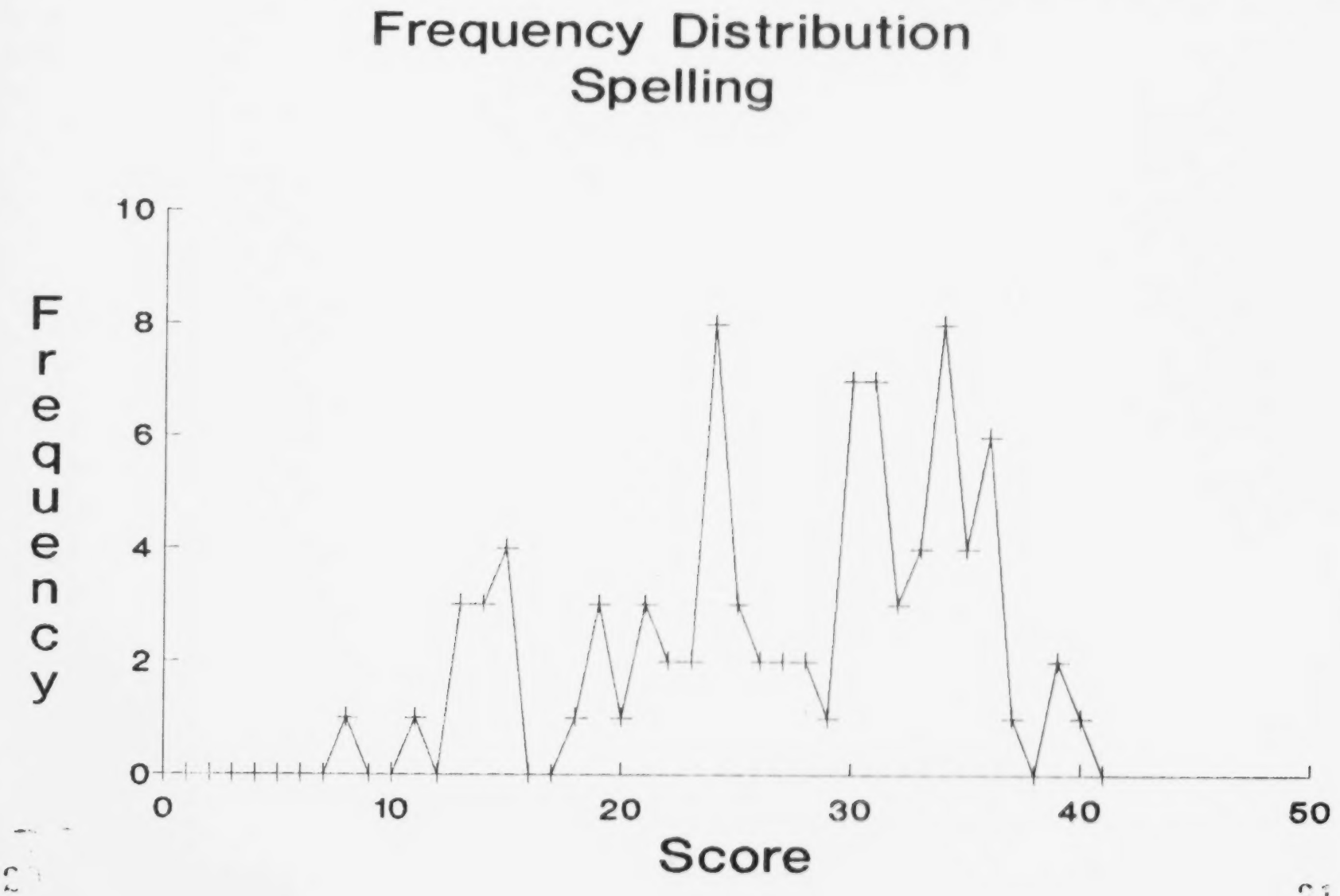
Frequency Distribution
Spelling
Frequency
10
8
6
4
2
0
0
10
20
30
40
50
Score

Figure 2

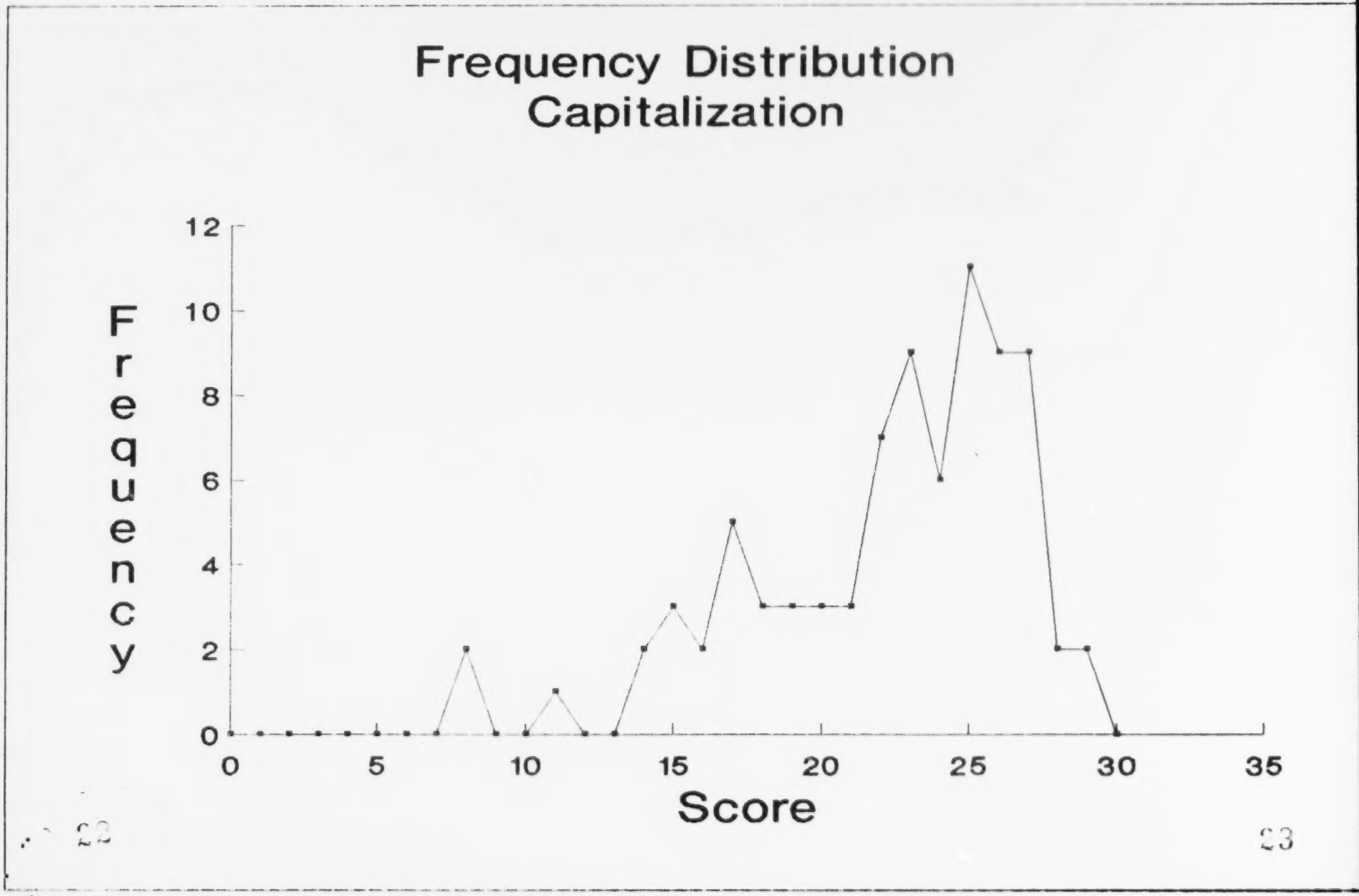
Frequency Distribution
Capitalization
Frequency
12
10
8
6
4
2
0
0
5
10
15
20
25
30
35
Score

Figure 3

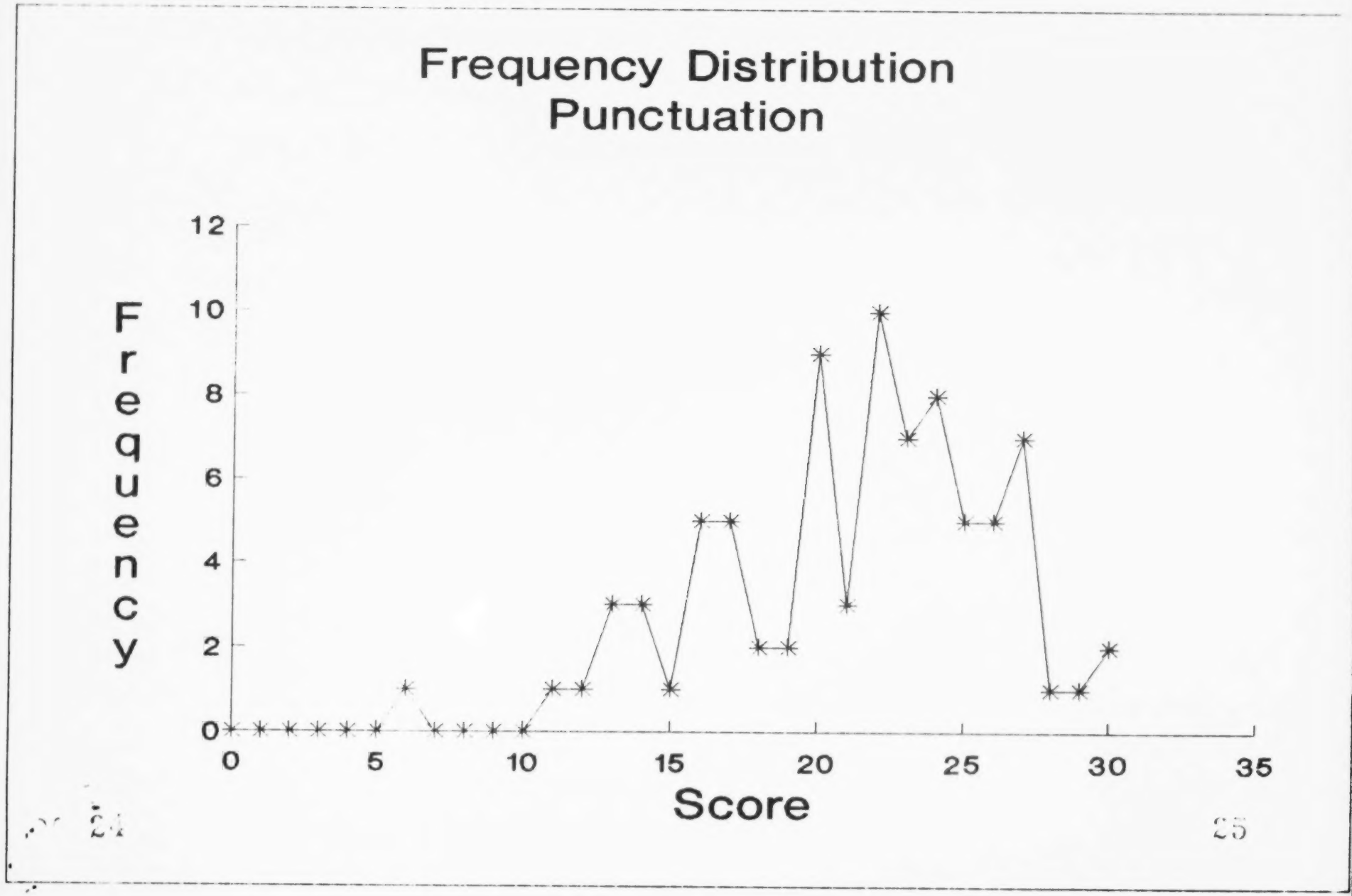
Frequency Distribution
Punctuation
Frequency
12
10
8
6
4
2
0
0
5
10
15
20
25
30
35
Score

Figure 4

Frequency Distribution Usage and Expression